نشریه پژوهشهای علوم دامی ایران
جلد ۱۲، شماره ۱، بهار ۱۳۹۹، ص. ۱۲۴-۱۱۳
Iranian Journal of Animal Science Research
Vol. 12, No. 1, Spring 2020, p. 113-124

بهینه‌سازی پارامترهای روش‌های یادگیری ماشین بر ارزیابی ژنومی صفات گسسته دودویی با در نظر گرفتن ساختار جمعیت و توزیع‌های متفاوت فنوتیپ در جمعیت مرجع

یوسف نادری*

تاریخ دریافت : ۱۳۹۷/۱۱/۰۵
تاریخ پذیرش: ۱۳۹۸/۰۳/۲۰

چکیده

تنظیم اولیه و بهینه‌سازی پارامترهای ورودی روش‌های یادگیری ماشین گامی اساسی جهت دستیابی به حداکثر صحت پیش‌بینی ژنومی می‌باشد. در این تحقیق، جمعیت‌های ژنومی برای سطوح مختلف وراثت‌پذیری (۰/۰۵ و ۰/۲)، عدم تعادل پیوستگی (پایین و بالا) و تعداد متفاوت جایگاه صفات کمی (۲۰۰ و ۶۰۰) بر روی ۲۹ کروموزوم شبیه‌سازی شد. جهت ایجاد نسبت‌های مختلف فنوتیپ آستانه‌ای دودویی، فنوتیپ افراد جمعیت مرجع وابسته به اینکه باقی‌مانده آنها کمتر از ē-1SDe (رویکرد اول) یا ۵۰ درصد افراد جمعیت (رویکرد دوم) باشد کد یک (فنوتیپ نامطلوب) و سایر حیوانات کد صفر (فنوتیپ مطلوب) اختصاص داده شد. برای بهینه‌سازی پارامترهای ورودی مدل، سطوح مختلف تعداد SNP نمونه‌گیری‌شده (۱۰۰، ۱۰۰۰ و mtry=۲۰۰۰)، تعداد بوت استراپ (۵۰۰، ۱۰۰۰ و ntree=۲۰۰۰) و حداقل اندازه گره پایانی (۱ و node size=۵) برای جنگل تصادفی و سطوح مختلف تعداد درخت (۱۰۰، ۱۰۰۰ و ntree=۲۰۰۰)، عمق درخت (۱، ۵ و tc=۱۰) و نرخ یادگیری (۰/۱ و lc=۰/۰۵) برای Boosting در نظر گرفته شد. کمترین میزان خطای خارج از کیسه برای mtry برابر با ۲۰۰۰، ntree برابر با ۱۰۰۰ و node size برابر با ۱ و کمترین خطای اعتبارسنجی در روش Boosting برای ntree، tc و lr به ترتیب ۱۰۰۰، ۱۰ و ۰/۰۵ مشاهده شد. صحت پیش‌بینی ژنومی روش‌های جنگل تصادفی و Boosting با کاهش فنوتیپ نامطلوب (رویکرد اول) افزایش یافت. بطور کلی در تمام سناریوها روش Boosting عملکرد بهتری نسبت به روش جنگل تصادفی داشت که دلیل این امر را می‌توان لحاظ کردن اثرات متقابل بین نشانگرها، خود ترمیمی و قدرت بالای این روش در کاهش خطای مدل دانست.

واژه‌های کلیدی: اعتبارسنجی، صفات آستانه‌ای، عدم تعادل پیوستگی، وراثت‌پذیری، یادگیری ماشین

مقدمه

در سالهای اخیر، به کارگیری انتخاب ژنومی همراه با توسعه تکنولوژی فن‌آوری تعیین ژنوتیپ منجر به تسهیل پیشرفت ژنتیکی در برنامه‌های اصلاح‌نژادی شده است. در حقیقت صحت پیش‌بینی ژنومی از طریق انتخاب ژنومی افزایش و به سرعت در برنامه‌های اصلاح نژادی و خصوصا برای صفات کمی گسترش یافته است. از دیدگاه اصلاح نژادی پرداختن محض به این نوع صفات به علت همبستگی منفی با برخی صفات آستانه‌ای از جمله مقاومت به بیماری‌ها، درجه سختی زایش و صفات رفتاری منجر به کاهش شایستگی ژنتیکی حیوان خواهد شد (۸). در نتیجه، پیشرفت قابل توجه در افزایش سوددهی اقتصادی در برنامه‌های اصلاح نژادی مدرن نیازمند فهم بهتر و ورود مستقیم به صفات با بروز فنوتیپی گسسته دارد (۹، ۴۰) که این امر اهمیت پرداختن به این صفات را در تحقیق حاضر دو چندان می‌کند.

ماهیت آستانه‌ای این نوع صفات، اثر پذیری بوسیله ژن‌های متعدد و عدم تطابق با توزیع نرمال و وراثت مندلی چالش جدیدی را از منظر آماری در این نوع صفات مطرح کرده است (۳۶). در نتیجه، استفاده بهتر از انتخاب ژنومی و کاربرد آن در اصلاح‌نژاد دام به فهم بالای روش‌های آماری مورد استفاده در انتخاب ژنومی وابسته است. روش‌های آماری متعددی برای تخمین اثرات نشانگرها ارائه شده است. در دو دهه اخیر مویسن و همکاران (۲۲) با معرفی برخی الگوهای آماری انقلاب شگرفی در جهت ارزیابی ژنومی ایجاد نمودند. با این حال در سال‌های اخیر روش‌های یادگیری ماشین (۵) به طور گسترده‌ای جهت حل چالش‌های ارزیابی ژنومی صفات آستانه‌ای مطرح شده‌اند (۳۷). جنگل تصادفی (۲۴) و Boosting (۱۱) از جمله

استادیار، گروه علوم دامی، باشگاه پژوهشگران جوان و نخبگان، دانشگاه آزاد اسلامی، واحد آستارا، آستارا، ایران
(*-ایمیل نویسنده مسئول: yousefnaderi@gmail.com)
DOI:10.22067/ijasr.v12i1.78810

روش‌های قدرتمند یادگیری ماشین هستند که علاوه بر قدرت بالا در برآورد ارزیابی ژنومی (۱۱، ۱۴)، در تشخیص ژن–ژن، پروتئین–پروتئین، اثر متقابل ژن–محیط، ژن‌های مرتبط با بیماری، مدل‌سازی جهت ارتباط میان ترکیب نشانگرها، انتخاب ژن‌های در ارتباط با صفت هدف، شناسایی فاکتورهای تنظیمی در توالی آمینو اسیدها و DNA نقش برجسته‌ای دارند (۴۱). تحقیقات در مورد استفاده از روش‌های یادگیری ماشین در ارزیابی ژنومی نشان از برتری این روش‌ها (Boosting و جنگل تصادفی) نسبت به روش‌های بیز داشت (۱۴). مطالعات دیگر از قابل مقایسه بودن قدرت ارزیابی ژنومی جنگل تصادفی (۲۵) و Boosting (۱۱) در مقایسه با GBLUP سخن به میان آورده‌اند. با این حال، تفاوت عمده روش‌های یادگیری ماشین در مقایسه با روش‌ها مرسوم عدم نیاز به نحوه توارث، توانایی بالا در به کارگیری اثرات غیرافزایشی، فرضیات در نظرگرفته شده برای مدل ژنتیکی پشت صحنه آن‌ها و تنظیم و بهینه‌سازی پارامترهای اصلی آنها است که به عنوان مهمترین عوامل موثر جهت دستیابی به حداکثر صحت پیش‌بینی ژنومی شناخته می‌شود می‌باشد(۱۰، ۲۶).

علاوه بر تاثیر عواملی از جمله مدل آماری و نوع صفت مورد مطالعه، عوامل مختلفی می‌توانند صحت ارزش‌های اصلاحی ژنومی و ارزیابی ژنومی را تحت تأثیر قرار دهد، این عوامل شامل تعداد QTL (۳۹)، سطح عدم تعادل پیوستگی (۲۸، ۴۲)، تراکم نشانگرها (۳۸)، وراثت‌پذیری صفت (۲۷)، تعداد داده‌های فنوتیپی در جمعیت مرجع (۲۲) می‌باشند. با توجه به این که در انتخاب ژنومی ارزش‌های اصلاحی حیوانات جمعیت کاندیدا (دارای ژنوتیپ) از طریق برآورد میزان اثر هر کدام از نشانگرها بر صفت در جمعیت مرجع (دارای ژنوتیپ و فنوتیپ) برآورد می‌شود، با این حال برای صفات آستانه‌ای علاوه بر عوامل فوق الذکر، نسبت فنوتیپی جمعیت مرجع یکی از عوامل تاثیرگذار در برآورد ارزش‌های اصلاحی حیوانات جمعیت کاندیدا است (۳۰). تحقیقات در این زمینه نشان دادند که ترکیب جمعیت مرجع از عوامل مؤثر بر صحت پیش‌بینی ژنومی بوده (۲۱) و روش‌های یادگیری ماشین به شدت تحت تاثیر نرخ توزیع فنوتیپ جمعیت مرجع در برآورد صحت پیش‌بینی ژنومی قرار می‌گیرند (۲۴، ۲۵).

به طور کلی، بررسی جنبه‌های مختلف هر کدام از این عوامل بر صحت پیش‌بینی ژنومی علاوه بر هزینه‌بر بودن از نظر جنبه‌های اقتصادی، به صورت تجمعی و همزمان در داده‌های واقعی میسر نبوده، از این‌رو مطالعات شبیه‌سازی ابزاری ارزشمند برای ارزیابی و سنجش اعتبار روش‌های پیشنهادی در انتخاب ژنومی با هزینه‌ای خیلی کم می‌باشد که امکان پیش‌بینی تغییرات پارامترهای ژنتیکی بصورت همزمان را نیز فراهم می‌آورد. در این راستا و با توجه به تحقیقات محدود در زمینه تنظیم و بهینه‌سازی روش‌های بوستینگ و جنگل تصادفی و عدم مطالعه این فاکتور درارزیابی ژنومی صفات گسسته با توزیع متفاوت فنوتیپ آستانه‌ای، به عنوان یک نوآوری در تحقیق حاضر مورد بررسی قرار گرفت.

لذا پژوهش حاضر با هدف مطالعه اینکه آیا روش‌های ناپارامتری یادگیری ماشین مانند جنگل تصادفی و Boosting با در نظر گرفتن ساختار ژنتیکی و توزیع متفاوت فنوتیپ در جمعیت مرجع می‌توانند سیگنال‌های ژنتیکی را جهت رسیدن بهصحت پیش‌بینی ژنومی قابل قبولی میسر سازند یا خیر؟ در این مسیر اهمیت تنظیم و بهینه‌سازی فاکتورهای اولیه بر عملکرد هریک از این روش‌ها نیز به تفکیک مورد ارزیابی قرار گرفت.

مواد و روش‌ها

شبیه‌سازی سناریوها در تحقیق حاضر با استفاده از نرم افزار QMSim انجام گرفت (۳۴). نخست، برای شبیه‌سازی جمعیتی با عدم تعادل پیوستگی پایین، یک جمعیت پایه ۲۰۹۰ راسی برای ۱۰۰۰ نسل شبیه‌سازی شد. برای تولید جمعیتی با عدم تعادل پیوستگی بالاتعداد افراد جمعیت از طریق ایجاد یک گلوگاه ژنتیکی (Bottleneck) به ۲۰۹ راس در نسل ۱۱۰۰ کاهش یافت. در نهایت در آخرین جمعیت پایه، بعد از ۱۰۰ نسل (در نسل ۱۲۰۰) تعداد افراد جمعیت به ۲۰۹۰ راس افزایش داده شد. برای ایجاد جمعیت مرجع و تایید، همه افراد (۲۰۹۰ راس) آخرین نسل جمعیت پایه برای تولیدمثل در جمعیت حاضر مورد استفاده قرار گرفتند که در این بین ۶۰ راس نر در نظر گرفته شد تا منعکس کننده‌ی نسبت نر به ماده‌ی (حدود ۳ صدم و اندازه مؤثر جمعیت ۲۳۳) موجود در گله‌های گاو شیری باشد و بتوان اثر روش تلقیح مصنوعی بر نسبت نر به ماده را تقلید کرد (۴۲). نوع سیستم آمیزشی تصادفی بود و برای ۱۰ نسل دیگر (تا نسل ۱۲۱۰) جمعیت تکثیر شد. شانس آمیزش در همه‌ی حیوانات برابر (در هر دو جنس) و یک فرزند برای هر زایش در نظر گرفته شد. درصد جایگزینی برای نر و ماده به ترتیب ۵۰ و ۲۰ درصد در نظر گرفته شد. در جمعیت اخیر، انتخاب حیوانات برتر برای نسل بعد بر اساس ارزش اصلاحی بالا و معیار حذف بر اساس ارزش اصلاحی پایین و سن صورت گرفت. نشانگرها به صورت دو آللی و به صورت تصادفی برای هر یک از کروموزوم‌ها (۲۹ کروموزوم در دامنه‌ی ۴۲ تا ۱۵۸ سانتی مورگان) توزیع شدند. به ازای هر کروموزوم تعداد متفاوت نشانگر (دامنه‌ی ۱۸۰ تا ۶۵۳ با توجه به نوع کروموزوم) جهت تولید پنل‌های ۱۰ K شبیه‌سازی شد. در مجموع دو سطح مختلف ۲۰۰ QTL (با دامنه QTL ۴–۱۳ با توجه به نوع کروموزوم) و ۶۰۰ QTL (دامنه QTL ۹–۳۲ با توجه به نوع کروموزوم) در طول کروموزوم‌ها توزیع شدند. نرخ جهش برای نشانگرها و QTL ها در هر جایگاه و در هر نسل $۲/۵\times۱۰^{-۵}$ فرض شد (۳۰). دو سطح مختلف وراثت‌پذیری (۰/۰۵ و ۰/۲) برای هر صفت در نظر گرفته شد. در طراحی جمعیت

نهایی، افراد آخرین نسل (نسل ۱۲۱۰) به عنوان جمعیت تایید (۲۰۹۰ راس) در نظر گرفته شد که این افراد اطلاعات ژنوتیپی داشته اما فاقد اطلاعات فنوتیپی بودند. همچنین افراد ۴ نسل ما قبل جمعیت تایید (نسل ۱۲۰۶ تا ۱۲۰۹) در گروه جمعیت‌های مرجع (۸۳۶۰ راس) که این افراد هم اطلاعات ژنوتیپی داشته و هم ارزش‌های اصلاحی ژنومی آنها مشخص می‌باشد طبقه‌بندی شدند. با توجه به این که از دیدگاه آماری توزیع احتمال QTLهای صفات مهم اقتصادی توسط شمار اندک ژن‌های دارای اثر عمده و درصد بالایی از ژن‌ها کوچک اثر هستند و این فرضیه به توزیع گاما نزدیکتر است (۱۷). توزیع احتمال QTLها، گاما فرض شد. همچنین فراوانی آللی اولیه برای نشانگرها ۰/۵ در نظر گرفته شد. در هر نسل و هر جایگاه کل میزان واریانس افزایشی توسط QTL توجیه شد. در مجموع ۴ سناریو (سناریو اوّل: وراثت‌پذیری صفت ۰/۲، ۶۰۰QTL و عدم تعادل پیوستگی ۰/۲۰۵ (در فاصله ۰/۰۵ سانتی مورگان) ؛ سناریو دوّم: وراثت‌پذیری صفت ۰/۲، QTL ۲۰۰ و عدم تعادل پیوستگی ۰/۲۰۵؛ سناریو سوّم: وراثت‌پذیری صفت ۰/۰۵، QTL ۲۰۰ و عدم تعادل پیوستگی ۰/۲۰۵؛ جمعیت چهارم: وراثت‌پذیری صفت ۰/۰۵، تعداد QTL ۲۰۰ و عدم تعادل پیوستگی۰/۳۲۱) در تحقیق حاضر شبیه‌سازی شد.

برای ایجاد نسبت‌های مختلف فنوتیپ آستانه‌ای دودویی تغییراتی در فایل فنوتیپ خروجی QMSim ایجاد شد. به طوریکه فنوتیپ پیوسته حیوانات به عنوان متغیر پاسخ (y) از طریق عوامل ثابت یا مستقل (اثر نسل =x1 و اثر جنسیت =x2) آنالیز واریانس شد تا اثر عوامل ثابت و تاثیرگذار بر فنوتیپ تصحیح شود و در نهایت مقادیر باقی‌مانده برای هر حیوان محاسبه شدند. در نتیجه برای شبیه‌سازی فنوتیپ آستانه‌ای دودویی در جمعیت مرجع، ابتدا باقی‌مانده‌ها از بیشترین به کمترین مرتب شدند. در مرحله بعد با توجه نرخ توزیع فنوتیپ، باقی‌مانده‌های پیوسته از طریق دو رویکرد به فنوتیپ آستانه‌ای تبدیل شدند. در رویکرد اول: حیواناتی از جمعیت مرجع که مقادیر باقی‌مانده‌ی آنها از $\bar{e}$-1SD$_e$ کمتر بود، کد یک (یا فنوتیپ غیر مطلوب: حدود ۱۶ درصد) و سایر حیوانات کد صفر (یا فنوتیپ مطلوب: حدود ۸۴ درصد) اختصاص داده شد. در رویکرد دوم: تعدادی (۳۴ درصد تعداد اولیه) از حیوانات سالم رویکرد اول بطور تصادفی کد ۱ و بیمار در نظر گرفته شدند تا نسبت حیوانات سالم به بیمار به نسبت برابر ۵۰:۵۰ اختصاص داده شد. در این تحقیق نشانگرهای با فراوانی آللی کمیاب کمتر از ۰/۰۱ حذف شدند. برای ارزیابی صحت مدل‌ها ۱۰ تکرار برای هر سناریو در نظر گرفته شد.

سطح عدم تعادل پیوستگی برای سناریوهای مختلف شبیه سازی‌شده با استفاده از محاسبه‌ی توان دوم ضریب همبستگی (r^2) بین همه‌ی جفت نشانگرهای ممکن ارزیابی گردد (۱۸).

$$r^2=D^2/f(A)f(a)f(B)f(b)$$

در این فرمول، $D=f(AB)-f(A)(B)$ بوده و $f(AB)$، $f(A)$، $f(a)$، $f(B)$ و $f(b)$ به ترتیب، فراوانی‌های مشاهده شده هاپلوتایپ AB و آلل‌های A، a، B و b می‌باشند. نرم‌افزار PLINK 1.9 برای برآورد LD بین جفت نشانگرهای مختلف در ژنوم همه حیوانات موجود در آخرین نسل مورد استفاده قرار گرفت (۳۱).

مدل آماری

جنگل تصادفی

جنگل تصادفی با استفاده از نمونه‌برداری پیاپی از جمعیت و به دست آوردن تقریبی از توزیع واریانس صفت انجام می‌شود. در داده‌های اعتبار سنجی، درختان طبقه‌بندی توسط بوت استراپینگ در آنالیز جنگل تصادفی ساخته می‌شوند. از طریق استراتژی بگینگ و انتخاب متغیر تصادفی، جنگل تصادفی باعث کاهش خطای پیش‌بینی ژنومی می‌شود. مدل کلی جنگل تصادفی به صورت زیر است.

$$\hat{f}_{rf}^{P}(x)=\frac{1}{P}\sum_{p=1}^{P}T(x;\ \Psi p)$$

در اینجا Ψp، pامین درخت و برای هر مشاهده $\hat{f}_{rf}^{P}(x)$ از طریق میانگین پیش‌بینی‌های هر درخت محاسبه می‌شود. $T(x;\ \Psi p)_1^P$ دربرگیرنده مشاهدات خارج از درخت می‌باشد. سایر حیواناتی که جز این نمونه‌گیری نیستند به عنوان خارج از مجموعه شناخته شده و در اعتبارسنجی هر درخت گزینش می‌شوند (۲۶).

جنگل تصادفی از مجموعه‌ای از درختان و با استفاده از تعداد زیادی نمونه از اطلاعات افراد جمعیت مرجع ایجاد می‌شود. سپس مدل ایجاد شده در جمعیت مرجع بر جمعیت تأیید اعمال می‌شود. در ابتدا یکی از نمونه ها وارد هر گره از هر درخت شده و از این نمونه اطلاعات یک نشانگر برای تقسیم بندی افراد مورد استفاده قرار می‌گیرد و در نهایت افراد بر اساس اطلاعات ژنوتیپی خود برای نشانگر انتخاب شده دسته‌بندی می‌شوند. این عمل در گره‌های متوالی انجام می‌شود تا در نهایت به گره‌های پایانی رسیده که در آنها حداکثر یکنواختی وجود خواهد داشت (۱۰). داده‌های ژنومی شبیه‌سازی‌شده از طریق بسته‌ی RanFoG (۱۴) و نرم افزار R مورد آنالیز قرار گرفتند.

روش‌های جنگل تصادفی جهت دستیابی به حداکثر صحت پیش‌بینی ژنومی نیازمند بهینه‌سازی و تنظیم پارامترهای اصلی خود هستند سه پارامتر اصلی و مهمی که در جنگل تصادفی در مورد کلاسه‌بندی بایستی تنظیم شود عبارت اند از mtry، تعداد SNP نمونه برداری شده در هر بار نمونه گیری تصادفی، ntree یا تعداد بوت استرپ و یا تعداد درختانی که بایستی رشد کنند و معیاری برای انتخاب بهترین SNPبرای تقسیم شدن هر گره است، nodesize، حداقل اندازه گره پایانی و که نشان دهنده‌ی تعداد مشاهدات در هر شاخه درخت است (۲۶). در این تحقیق بهترین ترکیب این سه پارامتر

از طریق محاسبه پارامتر خطای خارج از کیسه (OBB error) و به عنوان یکی از اهداف این تحقیق محاسبه گردید. در ارتباط با دادههای گسسته مقدار پیش فرض برای mtry، ntree و nodesize به ترتیب ۱۰۰، ۵۰۰، ۱ بود که در تحقیق حاضر مقادیر ۱۰۰، ۱۰۰۰ و ۲۰۰۰ برای mtry، ۵۰۰ و ۱۰۰۰ و ۲۰۰۰ برای ntree و ۱ و ۵ برای nodesize لحاظ گردید و برای هر ترکیب مقدار پارامتر خطای خارج از کیسه محاسبه شد. در نهایت بهترین ترکیب از این سه پارامتر جهت آنالیز نهایی جنگل تصادفی مورد استفاده قرار گرفت.

Boosting

در الگوریتم Boosting توابع پایه مورد استفاده شامل یادگیرندههای ضعیف مانند درخت رگرسیونی میباشند. در این الگوریتم سعی میشود تعدادی یادگیر پایه ضعیف (یاد گیرندههای بهتر از حالت تصادفی) که مکمل همدیگر هستند تولید شود و از طریق آموزش با استفاده از یادگیرندههای قبلی، یاد گیرندههای جدید قویتری ایجاد شود. در این الگوریتم توابع پایه مانند درختان رگرسیونی به صورت سریالی هر یک روی باقیمانده درخت قبلی اضافه میشوند در نتیجه اشتباه دستهبندی در درخت قبلی باعث کاهش مقدار خطا در درخت بعدی میشود (۱۱). این الگوریتم تا زمانی ادامه می یابد که خطای آخرین درخت به حداقل برسد. مدل کلی الگوریتم Boosting به صورت زیر میباشد.

$$f(x) = \sum_{m=1}^{m} \beta_m \, b(x;\gamma_m)$$

در این جا β_m (m=۱،۲،...، M) ضرائب توزیع و $b(x;\gamma_m)$ توابع ساده با چند متغیره x به همراه یک سری از پارامترها y میباشد. در این روش درختان رگرسیونی به عنوان یادگیرندههای پایه استفاده میشود. همچنین بهعلت توزیع دادههای گسسته از توزیع برنولی برای حداقل کردن تابع خطا استفاده شد (۱۱). نتیجهگیری نهایی از طریق میانگین تصحیح شده (به هر درخت یک وزنه داده میشود) مجموعه درختها انجام شد. این الگوریتم از طریق بستهی gbm در نرم افزار R مورد آنالیز قرار گرفت (۳۲).

در روش Boosting پارامترهای بهینهسازی شامل تعداد درخت (ntree)، عمق درخت (tree complexity) و پارامتر انقباضی یا نرخ یادگیری (learning rate) میباشند. در این تحقیق جهت دستیابی به بیشینه عملکرد و بهترین ترکیب این پارامترها مقدار خطای اعتبارسنجی (cross validation error) محاسبه شد. در ارتباط با دادههای گسسته مقدار پیش فرض برای ntree ، tc و lr به ترتیب ۱۰۰، ۱، ۰/۱ بود که در تحقیق حاضر مقادیر ۱۰۰ و ۱۰۰۰ و ۲۰۰۰ برای ntree؛ ۱، ۵ و ۱۰ برای tc؛ و ۰/۱ و ۰/۰۵ برای lr لحاظ گردید. در نهایت بهترین ترکیب از این سه پارامتر جهت آنالیز نهایی Boosting مورد استفاده قرار گرفت. در نهایت صحت پیشبینی ژنومی از طریق همبستگی پیرسون بین ارزشهای اصلاحی پیشبینیشده و ارزشهای اصلاحی واقعی محاسبه شد (۲۶).

نتایج

نتایج تنظیم و بهینهسازی پارامترهای مهم مدلهای جنگل تصادفی و Boosting در جدول ۱ ارائه شده است. نتایج برای مدل جنگل تصادفی نشان داد که کمترین میزان خطای OOB برای mtry برابر با ۲۰۰۰، ntree برابر با ۱۰۰۰ و nodesize برابر با ۱ ایجاد شده است. این در حالی بود که در Boosting جهت دستیابی به کمترین میزان خطایا اعتبارسنجی (CV) مقادیر بهینه ntree، tc و lr به ترتیب ۱۰۰۰، ۱۰ و ۰/۰۵ بودند.

پژوهشها در مورد مطالعه همخوانی سراسر ژنوم جهت بهینهسازی پارامترهای مهم جنگل تصادفی نشان داد که تعداد SNP نمونه برداری شده در هر بار نمونه گیری تصادفی (mtry) باید همواره بیش از ۱۰ درصد تعداد کل نشانگرها باشد (۱۳). این در حالی است که در مورد مدلهای پیوسته و گسسته مقدار mtry به ترتیب یک سوم تعداد نشانگرها و مجذور تعداد نشانگرها گزارش شده است (۵). با این حال هنگامی که دادهها دارای اطلاعات پرت باشند این مقادیر بیشتر در نظر گرفته میشوند، زیرا با افزایش تعداد mtry، دقت پیش بینی ژنومی در دادههای واقعی کمتر تحت تاثیر دادههای پرت قرار میگیرد (۴). این میزان بهعلت کاهش دادههای پرت در دادههای شبیهسازیشده کمتر به وقوع میپیوندد. همچنین مقدار mtry وابسته به دو مقدار ntree و nodesize متغییر خواهد بود. پارامتر تعداد درخت (ntree) رابطه تقریبا خطی با افزایش تعداد نشانگر دارد به طوری که با افزایش تعداد نشانگرها تعداد درخت افزایش مییابد. با افزایش تعداد درخت احتمال نمونهگیری نشانگرهایی که در ارزیابی ژنومی مشارکت دارند بیشتر شده و شانس نمونهگیری هر نشانگر برای حداقل یک بار نمونهگیری بالا میرود. در نتیجه همواره باید تعداد ntree بالا در نظر گرفته شود اما این افزایش (در پارامترهای mtry و ntree) تاوان محاسباتی بالایی را به همراه خواهد داشت. از دیدگاه دیگر کاهش تعداد درخت اگرچه در برخی موارد صحتی برابر با تعداد بالای درخت داشته باشد اما قابلیت اطمینان و تکرار پذیری آن پایین باشد (۴). همچنین همواره افزایش تعداد درخت تضمین کنندهی صحت بالا نبوده و تنظیم همزمان آن با دو پارامتر دیگر مهمترین اصل در بهینهسازی میباشد. حداقل تعداد مشاهدات در گره پایانی (nodesize) نشان میدهد چه تعداد حیوان در گره پایانی تجمع داشته باشند و افزایش این پارامتر منجر به تولید درختانی کوچک و کاهش زمان محاسباتی به دلیل عدم انشعاب گره برای نشانگرها میشود. در نتیجه در مطالعات ژنومی و خصوصاً صفات آستانهای این پارامتر پایینتر در نظر گرفته میشود (۱۱، ۲۶). تحقیقات در مورد بهینهسازی

پارامترهای جنگل تصادفی برای پنل ۱۰k نشان داد که مقدار بهینه این پارامترها ۱۰۰ برای mtry، ۲۰۰۰ برای ntree و ۱ برای nodesize در صفات آستانه‌ای (۲۶) و ۶۰۰۰ برای mtry، ۱۰۰۰ برای ntree و ۵ برای nodesize در صفات پیوسته بود (۱۱). نتایج مطالعات شبیه‌سازی نشان داد که بهینه‌سازی و دانش در مورد رابطه داخلی بین متغیرهای تنظیمی از مهمترین فاکتورهای مؤثر بر صحت پیش‌بینی ژنومی روش‌های یادگیری ماشین هستند (۱۴).

با توجه به اینکه در روش Boosting درختان به صورت سریالی روی باقی‌مانده درخت قبلی و با تاکید بر داده‌های به اشتباه دسته‌بندی رشد می‌کنند لذا با افزایش تعداد ntree قدرت مدل افزایش و خطای آن کاهش می‌یابد که نتایج تحقیق حاضر مؤکد این موضوع است. با این حال افزایش ntree هزینه سنگین محاسباتی را به همراه خواهد داشت. دو پارامتر tc و lr تعیین کننده تعداد ntree می‌باشند به طوری که افزایش هر کدام از این پارامترها نیاز به ntree بالا را کاهش می‌دهد. نرخ یادگیری (lr) میزان مشارکت هر درخت در مدل را نشان می‌دهد و مقدار آن با تعداد درخت رابطه عکس دارد (۱۱) و معمولا مقادیر پایین نرخ یادگیری به همراه تعداد بالای درخت مطلوبیت بیش تری در کاهش خطای CV دارد. مقادیر بالای نرخ یادگیری وقوع بیش‌برآورد را بالا برده و تعمیم پذیری مدل را کاهش می‌دهد. در نتیجه مدل نمی‌تواند به خوبی داده‌ها را دسته‌بندی کند. عمق درخت (tc) از مهمترین فاکتورهای Boosting جهت مطالعات ژنومی بوده و اثرات متقابل بین نشانگرها را کنترل می‌کند. در تحقیق حاضر همواره افزایش اثرات متقابل نشانگرها در شرایط ثبات دو فاکتور دیگر منجر به کاهش میزان خطای CV شده که با نتایج سایر محققین مطابقت داشت (۱۱). لازم به ذکراست که اهمیت این پارامتر در حجم بالای اطلاعات مشهودتر است. با افزایش تعداد tc برهم‌کنش بین نشانگرها در مدل بیشتر لحاظ می‌گردد و نیاز به ntree و lr کاهش می‌یابد. تحقیقات جهت بررسی ترکیب بهینه پارامترهای Boosting نشان داده شد که مقادیر ntree برابر با ۱۵۰۰، tc برابر با ۷ و lr برابر با ۰/۰۲ کمترین میزان خطای CV و بیشترین میزان صحت پیش‌بینی ژنومی را به همراه خواهند داشت (۱۱).

جدول ۱- مقادیر خطای OOB و CV برای ترکیب‌های مختلف پارامترهای تنظیمی در مدل‌های جنگل تصادفی و Boosting

Table 1- The OOB and CV error values for different combinations of tuning parameters in random forest and Boosting models

بوستینگ Boosting				جنگل تصادفی Random forest			
Ntree[1]	Tc[2]	Lr[3]	CV error[4]	Mtry[5]	Ntree	Nodesize[6]	OOB error[7]
100	1	0.1	119.481	100	500	1	269.41
100	1	0.05	118.37	100	500	5	269.58
100	5	0.1	115.45	100	1000	1	267.32
100	5	0.05	101.54	100	1000	5	268.28
100	10	0.1	108.92	100	2000	1	266.52
100	10	0.05	94.14	100	2000	5	266.98
1000	1	0.1	97.24	1000	500	1	258.40
1000	1	0.05	106.82	1000	500	5	264.37
1000	5	0.1	100.30	1000	1000	1	255.94
1000	5	0.05	79.23	1000	1000	5	262.56
1000	10	0.1	85.19	1000	2000	1	255.25
1000	10	0.05	71.11	1000	2000	5	260.45
2000	1	0.1	99.42	2000	500	1	257.33
2000	1	0.05	83.82	2000	500	5	261.21
2000	5	0.1	88.56	2000	1000	1	251.15
2000	5	0.05	73.42	2000	1000	5	259.34
2000	10	0.1	92.15	2000	2000	1	253.54
2000	10	0.05	71.32	2000	2000	5	260.22

[1] The number of trees to grow
[2] Tree depth or tree complexity
[3] Shrinkage rate or learning rate
[4] Cross validation error
[5] The number of SNP randomly selected at each tree node
[6] The minimum size of terminal nodes of trees
[7] Out-of-bag error

صحت پیش‌بینی ژنومی ثر نرخ توزیع فنوتیپ آستانه‌ای در جمعیت مرجع بر صحت ژنومی

جدول ۲ میانگین صحت پیش‌بینی ژنومی روش‌های Boosting و جنگل تصادفی برای نسبت‌های مختلف فنوتیپ آستانه‌ای جمعیت مرجع در هریک از جمعیت‌های شبیه‌سازی شده را نشان می‌دهد. به طور کلی در همه‌ی سناریوها با کاهش فنوتیپ نامطلوب (رویکرد اول) صحت پیش‌بینی ژنومی افزایش یافت. همچنین برای هر دو رویکرد همواره صحت پیش‌بینی ژنومی روش Boosting نسبت به روش جنگل تصادفی بیشتر بود. به طور کلی کمترین و بیشترین صحت پیش‌بینی ژنومی به ترتیب در روش جنگل تصادفی برای رویکرد دوم و روش Boosting برای رویکرد اول مشاهده شد.

جدول ۲- صحت پیش‌بینی ارزش‌های اصلاحی ژنومی (SD) با استفاده از روش جنگل تصادفی و Boosting برای نسبت‌های مختلف نرخ فنوتیپ آستانه‌ای در جمعیت مرجع

Table 2- Accuracy (SD) of estimated GEBVs usingof Random Forest (RF) and Boosting for different proportions of the threshold phenotypes rate in training set

مدل Method	نسبت‌های متفاوت فنوتیپ آستانه در جمعیت مرجع Different proportion of threshold phenotype in training set	سناریو Scenario			
		I[1]	II[2]	III[3]	IV[4]
بوستینگ Boosting	50:50	0.560^{a} (0.04)	0.526^{a} (0.04)	0.400^{fg} (0.04)	0.467^{bcd} (0.04)
	$\bar{e}$- $1SD_e$	0.580^{a} (0.05)	0.575^{a} (0.03)	0.435^{def} (0.04)	0.488^{bc} (0.04)
جنگل تصادفی Random forest	50:50	0.520^{ab} (0.03)	0.416^{efg} (0.03)	0.287^{i} (0.05)	0.382^{gh} (0.04)
	$\bar{e}$- $1SD_e$	0.570^{a} (0.04)	0.461^{cde} (0.04)	0.336^{hi} (0.03)	0.413^{efg} (0.04)

بین تمام سناریوها، مقادیر صحت با حروف لاتین غیر مشترک دارای اختلاف معنی‌دار هستند (P<۰/۰۵).

Amounts of accuracy without common superscript within all scenarios are significantly different ($P<0.05$).

[1] $h^2 = 0.2$, LD = low and 600 QTL

[2] $h^2 = 0.2$, LD = low and 200 QTL

[3] $h^2 = 0.05$, LD = low and 200 QTL

[4] $h^2 = 0.05$, LD = high and 200 QTL.

نتایج تحقیق حاضر نشان از برتری روش Boosting نسبت به جنگل تصادفی در نسبت های مختلف نرخ فنوتیپ آستانه‌ای در جمعیت مرجع برای سناریوهای مختلف ژنومی داشت. بطور کلی افزایش نسبت آستانه اثر منفی بر صحت پیش‌بینی ژنومی داشت. تحقیقات نشان داد که هرچقدر نرخ آستانه شبیه‌سازی‌شده به شیوع یا آستانه‌ای واقعی جامعه نزدیک‌تر باشد صحت پیش‌بینی ژنومی افزایش می‌یابد. در این راستا مطالعات نشان داد که افزایش شیوع یا آستانه نامطلوب از ۵ به ۲۵ درصد موجب افزایش صحت پیش‌بینی ژنومی و مقادیر بالاتر آن (نرخ شیوع آستانه نامطلوب) افت صحت پیش‌بینی ژنومی را به همراه خواهد داشت (۲۴) که تاییدی بر نتایج حاصل از این تحقیق دارد. از طرفی دیگر، در تحقیق حاضر نرخ آستانه صفت از طریق کد گذاری فنوتیپ پیوسته (باقیمانده) اعمال شد که این عمل به نوبه خود موجب افزایش خطای آستانه‌سازی در رویکرد دوم نسبت به رویکرد اول شد. دلیل این امر این است که برای رویکرد دوم، تعداد افراد بیشتری (۳۴ درصد جمعیت رویکرد اول) بدون در نظر گرفتن شایستگی‌شان و تنها با استفاده از انتخاب تصادفی دسته بندی می شوند که این امر منجر به افزایش بیشتر خطای دسته بندی در رویکرد دوم (نسبت به اول) و کاهش صحت پیش‌بینی ژنومی می‌شود.

تحقیقات محدودی در زمینه اثر نرخ فنوتیپ آستانه‌ای جمعیت مرجع بر صحت پیش‌بینی ژنومی انجام شده است که اکثر این تحقیقات یک نوع نرخ آستانه فنوتیپ مورد ارزیابی قرار گرفت. در یک مطالعه شبیه‌سازی (۱۴)، جمعیت ژنومی با ۲۵۰۰ حیوان برای یک صفت گسسته دودویی با نرخ آستانه ۵۰ درصد در جمعیت مرجع نشان داد از برتری مدل L_h–Boosting (۰/۴۱ –۰/۳۴) نسبت به جنگل تصادفی (۰/۳۲ –۰/۲۶) داشت. در مطالعات شبیه‌سازی و همچنین ارزیابی ژنومی صفات آستانه‌ای گاوهای هلشتاین آلمان صفات باروری و ناهنجاری‌های سم و ورم پستان کلینیکی نسبت‌های مختلف شیوع بیماری در جمعیت مرجع شبیه‌سازی شد و نشان داده شد که صحت پیش‌بینی ژنومی با افزایش نرخ آستانه در جمعیت مرجع (از ۵) به ۲۰ درصد افزایش و پس روندی نزولی داشت و عملکرد GBLUP بهتر از جنگل تصادفی بود (۲۴). نادری (۲۷) جمعیت‌های مختلف ژنومی را برای نرخ آستانه ۵۰ درصدی در جمعیت مرجع برای صفات آستانه‌ای شبیه‌سازی کرد و نشان داد که علیرغم صحت قابل قبول در برخی سناریوها، روش جنگل تصادفی نسبت به روش‌های بیزی بهتر نبود.

اثر تعداد QTL بر صحت پیش‌بینی ژنومی در جمعیت‌های با نرخ متفاوت فنوتیپ آستانه‌ای

برای ارزیابی اثر تعداد QTL بر صحت ژنومی، سناریو ۱ (۲۰۰QTL) و ۲ (۶۰۰QTL) برای نسبت‌های مختلف فنوتیپ آستانه‌ای با هم مقایسه شدند (جدول ۲). بیشترین میزان صحت پیش‌بینی ژنومی (۰/۵۸) برای جمعیت ۱ (تعداد زیاد QTL) با استفاده از Boosting در رویکرد اول و کمترین مقدار صحت (۰/۴۱۶) برای جمعیت ۱ (با تعداد کم QTL) با استفاده از جنگل تصادفی در رویکرد دوم مشاهده شد. علیرغم عملکرد بهتر روش Boosting نسبت به جنگل تصادفی، هر دو روش عملکرد بهتری در سطوح بالای QTL نسبت به سطوح پایین QTL نشان دادند. مطالعات نشان داد که اثر تعداد QTL بر صحت پیش‌بینی ژنومی شدیدا به نوع مدل آماری وابسته است (۳۳). برای مثال در رویکرد دوم، تغییرات تعداد QTL به ترتیب منجر به تغییرات ۰/۰۳۴ و ۰/۱۰۴ واحدی در صحت پیش‌بینی ژنومی روش های Boosting و جنگل تصادفی شد که نشان می‌دهد روش جنگل تصادفی نسبت به Boosting به تغییرات QTL حساسیت بالاتری دارد. مطالعات نشان داد افزایش تعداد QTL می‌تواند منجر به افزایش صحت پیش‌بینی ژنومی شود هنگامی که به طور موازی با افزایش تعداد QTLها، تعداد نشانگرها نیز افزایش یابد، شانش به دام انداختن اثرات QTLها افزایش می‌یابد (۱۶) . در روش‌های بازنمونه‌گیری از جمله Boosting و جنگل تصادفی، افزایش تعداد QTL، منجر به تولید عدم پیوستگی قوی بین برخی نشانگرها با QTLهای کنترل کننده‌ی صفت، نزدیک‌تر شدن فاصله نشانگرها با QTLها و افزایش شانس نمونه‌گیری شده، در نتیجه افزایش صحت پیش‌بینی ژنومی را به همراه دارد (۲۶)، که این اثر مثبت در نتایج تحقیق حاضر صادق بود. از منظر دیگر، عبدالهی و همکاران (۱) بیان کردند که در تعداد QTL پایین، احتمال شکل‌گیری توزیع اثرات ژنی کم بوده و توزیع آماری مورد نظر با تعداد ژن‌های بزرگ اثر و کوچک اثر به خوبی بیان و نمایان نمی‌شود که می تواند محتمل‌ترین دلیل برای نتایج به دست آمده باشد. هم چنین به خاطر اینکه توزیع اثرات نشانگرها در این تحقیق گاما بود و این مطلب در تحقیقات مختلف اذعان شده است که این توزیع در مقایسه با توزیع نرمال اثرات نشانگرها به پیش فرض‌های روش یادگیری ماشین سازگاری کمتری دارد در نتیجه در تعداد اندک QTL، روش های یادگیری ماشین عملکرد بالایی نخواهند داشت (۲۶).

غفوری کسبی (۱۱) در یک مطاله شبیه سازی نشان داد که روش جنگل تصادفی عملکرد بهتری در تعداد بالای QTL (۱۰۰۰) نسبت به تعداد پایین QTL (۱۰۰) برای صفات با وراثت‌پذیری پایین از خود نشان می‌دهند. ژانگ و همکاران (۴۳) بیان کردند که صحت پیش‌بینی ژنومی به آرامی با افزایش تعداد QTL افزایش می‌یابد. هم چنین گرگانی فیروزجاه و همکاران (۱۵) نشان دادند که با افزایش تعداد QTL از ۴۰۰ به ۶۰۰ میانگین صحت پیش‌بینی ژنومی از ۴۲/۷ به ۴۳/۳ افزایش می یابد. در مجموع گزارشات ضد و نقیض در مورد اثر بخشی تعداد QTL بر صحت پیش‌بینی ژنومی مطالعات مختلف را می توان به عوامل مختلفی (علاوه بر روش آماری) از جمله تعداد کروموزوم (۶)، اندازه مؤثر جمعیت (۲)، معماری ژنتیکی صفت (۲۷)، طبیعت افزایشی سناریوهای شبیه سازی شده و تعامل پیچیده بین ژن‌ها و مسیرهای بیولوژیکی در داده‌های واقعی مرتبط دانست (۲۶).

اثر وراثت‌پذیری بر صحت پیش‌بینی ژنومی در جمعیت‌های با نرخ متفاوت فنوتیپ آستانه‌ای

برای بررسی اثر وراثت‌پذیری بر صحت پیش‌بینی ژنومی در جمعیت‌های با نرخ متفاوت فنوتیپ آستانه‌ای، سناریو ۲ (h^2=۰/۲) و ۳ (h^2=۰/۰۵) مورد ارزیابی قرار گرفت. بیشترین میزان صحت پیش‌بینی ژنومی برای روش Boosting (۰/۵۷۵) در سناریو با وراثت‌پذیری بالا- رویکرد اول و کمترین میزان صحت پیش‌بینی ژنومی برای جنگل تصادفی (۰/۲۸۷) در سناریو با وراثت‌پذیری پایین – رویکرد دوم مشاهده شد. به طور کلی صحت پیش‌بینی ژنومی ناشی از روش های Boosting و جنگل تصادفی در هر دو رویکرد با افزایش وراثت پذیری افزایش یافت که این نتایج با تئوری بو و همکاران (۳) در مورد ارتباط مستقیم بین وراثت‌پذیری و صحت پیش‌بینی ارزش‌های اصلاحی ژنومی مطابق بود. در روش‌های رایج انتخاب، روند بهبود صفات با وراثت‌پذیری پایین به آرامی صورت می‌گیرد، به دلیل این که صحت به میزان زیادی به وراثت‌پذیری صفت بستگی دارد(۲۳). به طورکلی دلیل افزایش صحت پیش‌بینی ژنومی با افزایش وراثت‌پذیری را می‌توان این گونه عنوان کرد که هر چه وراثت‌پذیری صفت بیشتر باشد، فنوتیپ فرد به ارزش ژنتیکی فرد نزدیک‌تر بوده و اثر نشانگرها و به دنبال آن ارزش‌های اصلاحی ژنومی افراد به طورصحیح‌تر پیش‌بینی می‌شود (۱۷، ۲۷).

در سالهای اخیر مطالعات زیادی جهت بررسی اثر وراثت‌پذیری بر صحت پیش‌بینی ژنومی انجام شده است که نشان می‌دهد وراثت‌پذیری از مهمترین فاکتورهای موثر بر صحت پیش‌بینی ژنومی است (۲۷، ۲۸). در مطالعه‌ای شبیه‌سازی اثر مثبت وراثت‌پذیری بر صحت پیش‌بینی ژنومی روش‌های Boosting و جنگل تصادفی به اثبات رسید به طوری که افزایش وراثت‌پذیری از ۰/۱ به ۰/۵ افزایش ۷۲/۴ و ۷۵/۵ درصدی را در صحت روش‌های Boosting و جنگل تصادفی به همراه داشت (۱۱). با این حال در بررسی اثر سطوح مختلف وراثت‌پذیری بر صحت پیش‌بینی ژنومی جمعیت موش تفاوت محسوسی در صحت پیش‌بینی ژنومی روش‌های یادگیری ماشین از جمله جنگل تصادفی و SVM مشاهده نشد (۲۹). مطالعات دیتویلر و

همکاران (۷) در برآورد صحت پیش‌بینی ژنومی از طریق فرمول $r = \sqrt{N_p h^2 / N_p h^2 + M_e}$ (که در اینجا N_P: تعداد افراد جمعیت مرجع، M_e: تعداد سیگمنت‌های کروموزوم مستقل و h^2: وراثت‌پذیری) نشان داد که صحت پیش‌بینی ژنومی ارتباطی مستقیمی با وراثت‌پذیری دارد. در چندین مطالعه اثر مطلوب افزایش وراثت‌پذیری بر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی ناشی از مدل‌های یادگیری ماشین به اثبات رسیده است. این تاثیر مثبت و مطلوب وراثت‌پذیری منجر به تغییرات بالای ژنتیکی و در نتیجه کمک به پیش‌بینی بهتر اثرات نشانگری شد (۲۶، ۲۷).

اثر عدم تعادل پیوستگی بر صحت پیش‌بینی ژنومی در جمعیت‌های با نرخ متفاوت فنوتیپ آستانه‌ای

برای ارزیابی اثر عدم تعادل پیوستگی بر صحت پیش‌بینی ژنومی در جمعیت‌های با نرخ متفاوت فنوتیپ آستانه‌ای، سناریوهای ۳ (LD پایین) و ۴ (LD بالا) با هم مقایسه شدند (جدول ۲). میانگینLD برای جمعیت‌های ۳ و ۴ در فاصله ۰/۰۵ سانتی مورگان به ترتیب ۰/۲۰۵ و ۰/۳۲۱ بود. بیشترین میزان صحت پیش‌بینی ژنومی برای روش Boosting (۰/۴۸۸) در سناریو با عدم تعادل پیوستگی بالا- رویکرد اول و کمترین میزان صحت پیش‌بینی ژنومی برای جنگل تصادفی (۰/۲۸۷) در سناریو با عدم تعادل پیوستگی پایین – رویکرد دوم مشاهده شد. به طور کلی صحت پیش‌بینی ژنومی ناشی از روش های Boosting و جنگل تصادفی در هر دو رویکرد با افزایش سطح عدم تعادل پیوستگی افزایش یافت. سطح بالای LD میان QTLها و نشانگرها نشان می‌دهد نشانگرها سهم بالایی از واریانس ژنتیکی را به خود اختصاص می‌دهند (۱۲). در نتیجه هرچه میزان عدم تعادل پیوستگی بین QTLها و نشانگرها بیشتر باشد شانس قرار گرفتن نشانگرهای با عدم تعادل پیوستگی بالا در کنار هم برای روش‌های باز نمونه‌گیری بیشتر شده که این امر منجر به عملکرد مثبت و افزایش صحت پیش‌بینی ژنومی روش‌های Boosting و جنگل تصادفی می‌شود (۲۵، ۲۶). به عنوان یک اصل کلی، وجود LD بین نشانگرو QTL منبع اصلی اطلاعات است و نقش عمده‌ای در صحت پیش‌بینی ارزش‌های اصلاحی ژنومی ایفا می‌کنند (۲۰، ۳۵). جوناس و همکاران (۱۹) وجود LD در بین نشانگرها را عاملی تاثیرگذار در بهبود صحت پیش‌بینی ارزش‌های اصلاحی ژنومی عنوان کردند.

نتیجه گیری

تنظیم روش‌ها و استفاده از ترکیب بهینه‌ی پارامترهای روش‌های Boosting و جنگل تصادفی از ملزومات استفاده از این روش‌های در ارزیابی ژنومی می‌باشد. نرخ فنوتیپ در جمعیت مرجع یکی از مهم ترین عوامل مؤثر بر صحت پیش‌بینی ژنومی با استفاده از روش‌های Boosting و جنگل تصادفی بود. صحت پیش‌بینی ژنومی هنگام بیشینه بود که فنوتیپ نامطلوب صفت مورد نظر درصد کمی (۱۶ درصد) و نزدیک به جمعیت واقعی را داشته باشد. به طور کلی در تمام سناریوها روش Boosting عملکرد پیش‌بینی بهتری نسبت به روش جنگل تصادفی داشت که دلیل این امر را می‌توان لحاظ کردن اثرات متقابل بین نشانگرها، خود ترمیمی و تمرکز بالای این روش بر کاهش خطای مدل دانست به طوری که اشتباه دسته بندی در یک درخت توسط درختان دیگر به صورت رشد سریالی ترمیم و دسته‌بندی صحیح شده تا جایی که خطای مدل به حداقل برسد. علاوه بر نرخ آستانه صفت، نوع معماری ژنومی از عوامل مهم و مؤثر بر صحت پیش‌بینی ژنومی حاصل از روش‌های یادگیری ماشین بود. در این راستا اثر بخشی وراثت‌پذیری نسبت به تعداد QTL و میزان LD بر صحت پیش‌بینی ژنومی بیشتر بود. با وجود صحت پیش‌بینی ژنومی بالاتر روش Boosting در معماری های مختلف، هنگامی که صفات با وراثت‌پذیری بالا توسط تعداد زیادی QTL کنترل می شوند روش جنگل تصادفی برآورد قابل قبولی از صحت پیش‌بینی ژنومی ارائه داد.

منابع

1. Abdollahi-Arpanahi, R., A. Pakdel, A. Nejati-Javaremi, and M. M. Shahrbabak. 2013. Comparison of genomic evaluation methods in complex traits with different genetic architecture. Journal of Animal Production, 15:65-77.
2. Andonov, S., D. Lourenco, B. Fragomeni, Y. Masuda, I. Pocrnic, S. Tsuruta, and I. Misztal. 2017. Accuracy of breeding values in small genotyped populations using different sources of external information—A simulation study. Journal of Dairy Science, 100(1):395-401.
3. Bo, Z., J.-J. Zhang, N. Hong, G. Long, G. Peng, L.-Y. Xu, C. Yan, L.-P. Zhang, H.-J. Gao, and G. Xue. 2017. Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. Journal of Integrative Agriculture, 16(4):911-920.
4. Boulesteix, A. L., S. Janitza, J. Kruppa, and I. R. König. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6):493-507.
5. Breiman, L. 2001. Random forests. Machine Learning, 45(1):5-32.
6. Daetwyler, H., J. Hickey, J. Henshall, S. Dominik, B. Gredler, J. Van der Werf, and B. Hayes. 2010. Accuracy of

estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Animal Production Science, 50(12):1004-1010.

7. Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, andbenchmarking. Genetics, 193(2):347-365.
8. Egger-Danner, C., J. Cole, J. Pryce, N. Gengler, B. Heringstad, A. Bradley, and K. F. Stock. 2015. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits .Animal, 9(2):191-207.
9. Garrick, D. 2017. The role of genomics in pig improvement. Animal Production Science, 57(12):2360-2365.
10. Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar, and A. Nejati-Javaremi. 2016. Tuning and application of random forest algorithm in genomic evaluation. Research on Animal Production, 7(13):178-185 (In Persian).
11. Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar, and A. Nejati-Javaremi. 2017. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic BestLinear Unbiased Prediction in different scenarios of genomic evaluation. Animal Production Science, 57(2):229-236.
12. Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136(2):245-257.
13. Goldstein, B.A., A. E. Hubbard, A. Cutler, and L. F. Barcellos. 2010. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genetics, 11(1):49.
14. González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution, 43(1):7.
15. Gorgani Firozjah, N., H. Atashi, M. Dadpasand, and M. Zamiri. 2014. Effect of marker density and trait heritability on the accuracy of genomic prediction over three generations. Journal of Livestock Science and Technologies, 2(2):53-58.
16. Habier, D., R. L. Fernando, and J. C. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics, 182(1):343-353.
17. Hayes, B., and M. E. Goddard.2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution, 33(3):209.
18. Hill, W., and A. Robertson. 1968. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics, 38(6): 226-231.
19. Jónás, D., V. Ducrocq, and P. Croiseau. 2017. The combined use of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods enhances genomic evaluation accuracy in dairy cattle. Journal of Dairy Science, 10(4): 2905-2908.
20. Ke, X., S. Hunt, W. Tapper, R. Lawrence, G. Stavrides, J. Ghori, P. Whittaker, A. Collins, A. P. Morris, and D. Bentley. 2004. The impact of SNP density on fine-scale patterns of linkage disequilibrium. Human Molecular Genetics, 13(6):577-588.
21. Mc Hugh, N., T. Meuwissen, A. Cromie, and A. Sonesson. 2011. Use of female information in dairy cattle genomic breeding programs. Journal of Dairy Science, 94(8):4109-4118.
22. Meuwissen, T., B. Hayes, and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4):1819-1829.
23. Muir, W. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics, 124(6):342-355.
24. Naderi, S., M. Bohlouli, T. Yin, and S. König. 2018. Genomic breeding values, SNP effects and gene identification for disease traits in cow training sets. Animal Genetics, 49(3):178-192.
25. Naderi, S., T. Yin, and S. König. 2016. Random forest estimation of genomic breeding values for diseasesusceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. Journal of Dairy Science, 99(9):7261-7273.
26. Naderi, Y. 2018. Evaluation ofgenomic prediction accuracy in different genomic architectures of quantitative and threshold traits with the imputation of simulated genomic data using random forest method. Research on Animal Production, 9(20):129-138 (In Persian).
27. Naderi, Y. 2018. Impact of genotype imputation and different genomic architectures on the performance of random forest and threshold Bayes A methods for genomic prediction. Iranian Journal of Animal Science, 49(1):145-157 (In Persian).
28. Naderi, Y. 2018. Investigation of genotype× environment interaction with considering imputation in simulated genomic data via different animal models. Animal Production, 20(3):375-387 (In Persian).
29. Neves, H. H., R. Carvalheiro, and S. A. Queiroz. 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics, 13(1):100.
30. Pimentel, E. C., M. Wensch-Dorendorf, S. König, and H. H. Swalve. 2013. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genetics Selection Evolution, 45(1):12.

31. Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, and M. J. Daly. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3):559-575.
32. Ridgeway, G. 2017. Package 'gbm', the R project for statistical computing.
33. Sadeghi, S., S. A. Rafat, and S. Alijani. 2018. Evaluation of imputed genomic data in discrete traits using Random forest and Bayesian threshold methods. Acta Scientiarum. Animal Sciences, 40: e39007.
34. Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25(5):680-681.
35. Sun, X., R. Fernando, and J. Dekkers. 2016. Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. Genetics Selection Evolution, 48(1):77.
36. Wang, C., X. Ding, J. Wang, J. Liu, W. Fu, Z. Zhang, Z. Yin, and Q. Zhang. 2013. Bayesian methods for estimating GEBVs of threshold traits. Heredity, 110(3):213-219.
37. Wang, C., X .Li, R. Qian, G. Su, Q. Zhang, and X. Ding. 2017. Bayesian methods for jointly estimating genomic breeding values of one continuous and one threshold trait. PloS One, 12(4):e0175448.
38. Wang, Q., Y. Yu, J. Yuan, X. Zhang, H. Huang, F. Li, and J. Xiang. 2017 .Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei. BMC Genetics, 18(1):45.
39. Wientjes, Y. C., M. P. Calus, M. E. Goddard, and B. J. Hayes. 2015. Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genetics Selection Evolution, 47(1):42.
40. Yáñez, J. M., R. D. Houston, and S. Newman. 2014. Genetics and genomics of disease resistance in salmonid species. Frontiers in Genetics, 5:415.
41. Yang, P., Y. Hwa Yang, B. B Zhou, and A. Y Zomaya. 2010. A review of ensemble methods in bioinformatics. Current Bioinformatics, 5(4):296-308.
42. Yin, T., E. Pimentel, U. K. v. Borstel, and S. König. 2014. Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature humidity-dependent covariate. Journal of Dairy Science, 97(4):2444-2454.
43. Zhang, Z., Q. Zhang, and X. Ding. 2011. Advances in genomic selection in domestic animals. Chinese Science Bulletin, 56(26): 2655-2663.

Iranian Journal of Animal Science Research
Vol. 12, No. 1, Spring 2020, p. 113-124

نشریه پژوهشهای علوم دامی ایران
جلد ۱۲، شماره ۱، بهار ۱۳۹۹، ص. ۱۲۴-۱۱۳

The Effect of Parameters Tuning of Machine Learning Methods on Genomic Evaluation of Discrete Traits Considering Population Structure and Different Distributions of Phenotype in Training Set

Y. Naderi*

Received: 25-01-2019
Accepted: 10-06-2019

Introduction: The development of genotyping technologies has facilitated the genetic progress of breeding programs by implementing genomic selection (GS). In fact, the accuracy of genomic evaluations has been enhanced via GS and quickly spread in livestock breeding. For several decades, most phenotypic variation in dairy cattle populations had focused on continuous traits especially milk yield. From an animal breeding perspective, pay attention to this category of traits because of negative correlation with novel functional traits leads to reduction in genomic merit of these traits. Considerable advances along with increasing economic benefits in modern animal breeding programs requires better understanding and the direct inclusion of novel functional traits. Since many prominent traits in livestock including disease resistance and calving difficulty, present a binary distribution of phenotypes (and are often termed threshold traits), thus these traits are important in animal breeding due to importance of animal welfare and human tendency for healthy and high quality products. Threshold nature of most functional traits, affected by multiple genes, non-compliance from Mendelian inheritance and normal distribution are challenges for accurate prediction of GEBV using statistical methods in such kind of traits. Machine learning methodology as a non-parametric method commonly extended to solve the challenges of genomic selection for threshold traits. Random Forest (RF) and Boosting are powerful machine learning methods in order to recognize gene-gene, protein-protein and gene-environment interactions, to detect disease associated genes, to model the relationship among combinations of markers, to select genes associated with the target trait, to identify the regulatory factors in or protein and DNA sequences, to classify various samples in gene expression of microarrays data and to improve accuracy of genomic prediction. The objective of current study was to investigate the role of threshold phenotype rate of training set and different genomic architecture on performance of RF and Boosting methods. In this regard, per-determined and tuning input parameters of each method is a basic step to achieve maximum genomic accuracy.

Materials and Methods: A population of 2090 animals genotyped for 10,000 markers was simulated using QMSim software. In the first phase, over a time span of 1,000 generations, a historical population was provided from 1045 females and 1045 males. In the second phase, in order to produce a realistic level of LD, bottleneck was used. For this purpose, the population size decreased over 100 generations to 209 individuals. In the third phase, the population size increased over 100 generations (2030 females and 60 males). All 2090 individuals of the last historical generation served as founders and using a random mating design expanded the recent population by simulating an additional 10 generations. During these generations, replacement ratio was set at 0.2 and 0.50 for females and males, respectively and selection of candidate individuals were based on EBV and age. Each mating produced only one offspring with a same probability of being either male or female. Individuals of generations 6 to 9 was used as training set, while the whole generation 10 was considered as validation set. Genomic population were simulated to reflect variations in heritability (0.05 and 0.20), linkage disequilibrium (low and high) and number of QTL (200 and 600) for 29 chromosomes; therefore, four different scenarios including I (10K SNP, h^2 = 0.20, LD = low and 600 QTL), II (10K SNP, h^2 = 0.2, LD = low and 200 QTL), III (10K SNP, h^2 = 0.05, LD = low and 200 QTL) and IV (10K SNP, h^2 = 0.05, LD = high and 200 QTL) were simulated. In order to create different rates of discrete phenotype, the animals phenotype of training set was coded as1 (inappropriate phenotype) depending on whether their phenotype residuals was less than the average of residuals ($\bar{e}$) or $\bar{e}$- 1SD_e for the first and second approachs, respectively, and other individuals was defined ascode 0 (appropriate phenotype). In order to tuning input parameters of the model, different levels of *mtry*

Assistant Professor, Department of Animal Science, Young Resaerchers and Elite Club, Astara Branch, Islamic Azad University, Astara, Iran
(*- Corresponding Author Email: yousefnaderi@gmail.com)
DOI:10.22067/ijasr.v12i1.78810

(100, 1000 and 2000), *ntree* (500, 1000, 2000) and *nodesize* (1 and 5) for RF and *ntree* (500, 1000, 2000), *tc* (1, 5 and 10) and *lc* (0.1 and 0.05) for Boosting were considered.

Results and Discussion: The least of out-of-bag (OOB) error was obtained for *mtry*= 2000, *ntree*= 1000 and *nodesize*= 1 in RF method while the least of cross validation (CV) error was observed for boosting method with *mtry*= 2000, *tc*= 10 and *lc*= 0.05. In all scenarios, RF algorithm was showed a wide range of genomic accuracy (0.287 to 0.57) compared to Boosting method (0.4 to 0.58). Accuracy of genomic predicted was decreased in RF and Boosting with increasing the inappropriate phenotype, because of more individuals are in the vicinity of the average normal population for the first approach ($\bar{e}$)compared to the second approach ($\bar{e}$- 1SD_e), therefore leads to more classification errors (coding)and decrease of the genomic prediction accuracy. RF and Boosting showed a high performance when high-heritability traits were controlled by a large number of QTLs. Increase in number of QTLs generally led to a major improvement in RF accuracies, while a negligible positive effects were found for Boosting.

Conclusion: The composition of training set and population genomic architecture were two basic factors affecting accuracy of genomic prediction in machine learning methods. Interactions among predictive variables (SNP), self-healing and high potency to decrease training error were considered in Boosting method resulting in more accurate estimation in this method compared to the other RF method under all scenarios.

Keywords: Cross validation, Heritability, Linkage disequlibrium, Machine learning, Threshold traits